

SPECIFICATION

Electronic Version 1.2.8

Stylesheet Version 1.0

[PROCESSING DRUG DATA]

Cross Reference to Related Applications

This application is a continuation-in-part and claims priority to co-pending U.S. utility patent application No. 09/681,587, entitled Pharmacovigilance Database, filed May 02, 2001; and incorporates the disclosure of that application by reference in its entirety.

Background of Invention

- [0001] *FIELD OF THE INVENTION.* The present invention relates generally to systems and methods for processing drug information. More specifically, it relates to extracting data from drug information sources in a manner to support use of the data with artificial intelligence tools.
- [0002] *BACKGROUND.* Over 9,500 prescription drug products have been approved by the U.S. Food and Drug Administration (FDA). Label data for each drug is prepared by the drug manufacturer and approved by the FDA. Navigating through label data to locate information relevant to a prescribing decision, e.g., appropriate selection, dosing, cross-drug effects, contraindications, and warnings, is a daunting task for physicians, pharmacists, pharmaceutical benefit managers, hospital formularies, insurance companies, and others.
- [0003] Compilations of label data are available. The Physicians' Desk Reference ® (PDR) compiles full-length entries of the exact copy of most drug's FDA-approved label in hard copy. Computer-searchable versions of this data are available from the publisher of the PDR ® ; while computer-searchable versions of similar data are available from vendors such as Multum Information Services, Inc. Denver, Colorado and ePocrates, Inc., San Carlos, California.

[0004] Other drug information sources are available, such as articles from medical journals and formularies used by insurance carriers and health maintenance organizations (HMOs).

[0005] Each of these drug information sources may contain explicit and implicit information. For example, the drug label for RUBEX ® doxorubicin hydrochloride for injection includes the following adverse event content in text form:

[0006] *ADVERSE REACTIONS ... Cutaneous:* Reversible complete alopecia occurs in most cases. ... *Gastrointestinal:* Acute nausea and vomiting occurs frequently and may be severe.

[0007] The adverse event content above contains implicit information regarding an adverse event, e.g., *alopecia*, and its frequency of occurrence when the drug is used, i.e., *most*.

[0008] As a further example, the drug label for REMICADE™ infliximab includes the following adverse event data content in table form:

[0009]

| ADVERSE REACTIONS IN CROHN'S DISEASE TRAILS | | |
|---|---------------------|-------------------------|
| | Placebo (n = 56) | Infliximab (n = 199) |
| ... | | |
| Pts with ≥ 1 AE | 35 (62.5%) | 168 (84.4%) |
| WHOART preferred term | | |
| Headache | 12 (21.4%) | 45 (22.6%) |
| ... | | |

[0010] As another example, consider the drug label for PROZAC fluoxetine hydrochloride. Label adverse reaction information is given both explicitly in tables that contain percentages, and implicitly by use of the words *frequent*, *infrequent*, and *rare*.

[0011] In addition to adverse event data content, drug information sources, such as labels, typically contain instances of drug rule content. Instances of drug rule content include prose containing one or more drug rules. As an example consider the drug label for ENBREL® etanercept. Its label contains the following drug rule content

[0012] *CONTRAINDICATIONS*

[0013] ENBREL should not be administered to patients with sepsis or with known hypersensitivity to ENBREL or any of its compounds. ...

[0014] Typical existing approaches to managing drug information present the information in a simple manner, e.g., in a "warehouse" fashion, and do not focus on indirect or implicit information (especially adverse event data and drug rules). More specifically, existing approaches do not focus on capturing drug information in a manner amenable to use with artificial intelligence tools. Existing approaches typically focus on categorizing verbatim text without regard to the underlying logical content.

[0015] In addition, differing terminology employed by data authors also makes conventional queries cumbersome and the results less reliable than desired. This problem is acute in the area of medical information related to substances such as drugs. Drugs and other therapeutic substances may be known by a variety of names. In addition to the chemical name, many drugs have several clinical names recognized by health care professionals in the field. It is not uncommon for a drug to have several different trade names depending on the manufacturer. This matter is further complicated by one or more functional names that may be associated with a drug or other substance. For example, an antidepressant may be identified as Prozac[®], a fluoxetine, a serotonin reuptake inhibitor, or a serotonin receptor specific modulator. However, antidepressants include many other drugs, such as lithium and other catecholaminergic drugs, and there are serotonin reuptake inhibitors in addition to Prozac[®]. Even "standardized" terminology can differ between compilations. For example, references that can serve as sources of standard terminology include Medical Dictionary for Regulatory Activities (MedDRA[™]), World Health Organization Adverse Reaction Terminology (WHO-ART), or Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART) developed and maintained by the FDA's Center for Drug Evaluation and Research.

[0016] Typically, compilers of drug label data make minimal, if any, effort to improve the quality of the data. Data corruption can include extraneous non-alpha characters, noise words, misspellings, and dislocations (e.g., data that is valid for one category, erroneously entered into another, inappropriate field).

[0017] In addition, existing methods of compiling and organizing such data do not focus on the rules regarding drug safety contained within drug information sources. Existing approaches typically focus on categorizing verbatim text without regard to the underlying logical content.

[0018] Existing methods, alone or in combination, do not address improving the quality of the underlying verbatim drug information source data. Nor do existing methods address mapping this underlying data to accepted pharmaceutical community terms and hierarchies through which to direct queries. The problem of differing terminology among the disparate labels also remains un-addressed; as does the problem of data corruption in the form of misspelling and extraneous characters.

[0019] Typical existing methods of processing drug information are not focused on extracting rules or adverse event data from drug information sources. Nor do those methods address structuring these rules in a format amenable to use by inference engines, reasoning engines, or other similar sophisticated data processing techniques.

[0020] In view of the above-described deficiencies associated with data concerning drugs and other substances associated with medical databases, there is a need to solve these problems and enhance the amenability to efficient use of such data. These enhancements and benefits are described in detail herein below with respect to several alternative embodiments of the present invention.

Summary of Invention

[0021] The present invention in its disclosed embodiments alleviates the drawbacks described above with respect to existing drug information databases and incorporates several additionally beneficial features.

[0022] In some embodiments, the invention includes computer-assisted methods of processing a drug information source into syntax-parsed drug rules. The method includes creating a drug rule syntax; detecting drug rule content from the drug information source; and parsing drug rule elements from drug rule content into the drug rule syntax. In those embodiments, the associations between those drug rule elements that form a drug rule are retained.

[0023] In other embodiments, the invention includes computer-assisted methods of processing a drug information source into adverse event characterizations. In those embodiments, the method includes: detecting adverse event content from the drug information source and parsing adverse event characterizations.

[0024] In further embodiments, the invention includes computer-assisted methods for processing a drug information source to characterize the drug by the set comprising: syntax-parsed drug rule elements, adverse event data, mapped terms, and metadata. In those embodiments, the method includes: creating a drug rule syntax; extracting metadata from the drug information source; extracting verbatim adverse event data from the drug information source; identifying drug rule content from the drug information source; mapping terms from verbatim data to a reference source; and parsing drug rule elements from at least one identified instance of drug rule content into the drug rule syntax, retaining associations between those drug rule elements that form a drug rule.

[0025] It is an object of the present invention to rationalize drug information source data into a structure amenable to efficient query. In addition, a feature of preferred embodiments of the invention is that processing data in a fashion of the invention permits more than just effective database query. It permits operations to be performed on the data, e.g., calculations, comparisons, rule triggering. For example, forward and backward chaining inference engines require a rules base, Fuzzy logic requires a probabilistic or lexical way to assess closeness. Neural networks require taxonomies that allow for propagation of information through the network. Analogical reasoning, or case-based reasoning, requires a format to describe *stories* or situations whose relevance can be calculated using known techniques.

[0026] It is an object of the present invention to develop a drug database amenable to query using canonical terms accepted in the pharmaceutical industry. Linking terms to standard vocabulary for data such as drug name and reaction enables meaningful statistical comparisons to be made.

[0027] The beneficial effects described above apply generally to the exemplary systems and methods for developing a drug database. The specific structures through which these benefits are delivered will be described in detail hereinbelow.

Brief Description of Drawings

- [0028] The invention will now be described in detail, by way of example without limitation thereto and with reference to the attached figures.
- [0029] Figure 1 is a conceptual relationship diagram of preferred embodiments of the present invention for processing drug rules.
- [0030] Figure 2 is an example of a rule structure.
- [0031] Figure 3 is an example of parsing detected instances of rule content as drug rule elements into a syntax.
- [0032] Figure 4 is an illustrative data flow diagram of preferred embodiments of the present invention.
- [0033] Figure 5 is an illustrative data flow diagram of preferred embodiments of the present invention.
- [0034] Figure 6 is an example of parsing detected instances of rule content as drug rule elements into a syntax.
- [0035] Figure 7 is an illustrative data flow diagram of preferred embodiments of the present invention.
- [0036] Figure 8 is an illustrative data flow diagram of preferred embodiments of the present invention.
- [0037] Figure 9 is a conceptual relationship diagram of preferred embodiments of the present invention for processing adverse event data.
- [0038] Figure 10 is an illustrative data flow diagram of preferred embodiments of the present invention.

Detailed Description

- [0039] As required, detailed embodiments of the present invention are disclosed herein; however, it is to be understood that the disclosed embodiments are merely exemplary of the invention that may be embodied in various and alternative forms. The figures are not necessarily to scale, some features may be exaggerated or minimized to show

details of particular components. Therefore, specific structural and functional details disclosed herein are not to be interpreted as limiting, but merely as a basis for the claims and as a representative basis for teaching one skilled in the art to variously employ the present invention.

[0040] In a preferred embodiment, the present invention includes a computer-assisted method of processing a drug information source. As illustrated in Figure 1, a drug information source 10 typically includes at least one instance of drug rule content 20. For example, the drug label (a drug information source 10 for ENBREL[®] brand etanercept includes the following drug rule content 20.

[0041] " *CONTRAINDICATIONS*

[0042] ENBREL should not be administered to patients with sepsis or with known hypersensitivity to ENBREL or any of its components."

[0043] Each instance of drug rule content 20 includes one or more specific drug rules 30. The immediately prior example includes the following drug rule.

[0044] "ENBREL should not be administered to patients with sepsis"

[0045] In preferred embodiments, a drug rule syntax 40 is established. The drug rule syntax 40 comprises types of subsets of logical propositions, i.e., drug rule syntax elements 42, along with the allowable relationships between the types. The drug rule syntax 40 provides structure amenable to use by artificial intelligence engines. In the present example, the type *<drug in question>* can be instantiated as *ENBREL[®]*, the type *<concurrent condition>* can be instantiated as *sepsis*, and the type *<prescribing action for drug in question>* can be instantiated as *should not be administered*. For this example, a syntax rule may be stated as a proposition constructed as:

[0046] For every patient *p*, if *p* (has *<concurrent condition>* and *<drug in question>* is considered), then *<drug in question>* *<prescribing action>*.

[0047] Figure 2 illustrates another example of a rule structure.

[0048] Returning to Figure 1, given a drug information source 10 (e.g., a drug label, a

medical journal article, a formulary) and a syntax 40, preferred methods of the present invention include detecting at least one instance of drug rule content 20 from a drug information source 10. This step can be accomplished manually or interactively in a computer-assisted manner. Natural language processing (NLP) is suited to providing computer assistance at this step. NLP processes keying on phrases such as "should not be administered" or keying on all sentences under a heading CONTRAINDICATION can suggest sections of text as candidate drug rule content.

[0049] In preferred embodiments, instances of drug rule elements 44 are parsed from detected instances of drug rule content 20 into the drug rule syntax 40. The associations between those instances of drug rule elements 44 that form an instance of a drug rule 46 are retained. Using the immediately prior example for ENBREL[®], the set of drug rule elements { <drug in question> = ENBREL, <concurrent condition> = sepsis, <prescribing action> = should not be administered} is saved and associated with a proposition of the type constructed above. Further examples of parsing from drug rule content 20 to a drug rule syntax are illustrated in Figures 3 – 8. Figure 3 illustrates a complete parsing, into syntax elements, of the drug rule content:

[0050] " *CONTRAINDICATIONS*

[0051] ENBREL should not be administered to patients with sepsis or with known hypersensitivity to ENBREL or any of its components."

[0052] Figure 4 illustrates a complete parsing of the drug rule content:

[0053] *WARNINGS ...*

[0054] *PATIENTS WHO DEVELOP A NEW INFECTION WHILE UNDERGOING TREATMENT WITH ENBREL SHOULD BE MONITORED CLOSELY. ADMINISTRATION OF ENBREL SHOULD BE DISCONTINUED IF A PATIENT DEVELOPS SERIOUS INFECTION OR SEPSIS. ...*

[0055] Figure 5 illustrates a method of preferred embodiments of the present invention for processing rule content from a drug information source. In this example, either a user or a natural language processor detects words that are related to the standardized terms associated with the syntax elements. These terms or phrases are then mapped parsed into appropriate elements of the syntax. If an extracted term

matches a standardized term, then it is used. If not, then it is linked to a standardized term. The mapping between the verbatim text and the structured rule is retained as a pedigree for treacability.

[0056] Figures 6, 7, and 8 show additional examples of parsing. Figure 8 uses a horizontal mapping in a simple spreadsheet to show parsing without the use of natural language processing.

[0057] In some embodiments, the invention includes a computer-assisted method of processing a drug information source for extracting adverse event characterizations in a format amenable to use with artificial intelligence engines. Referring to Figure 9, the drug information source 210 comprising at least one instance of adverse event content 220. For example, the drug label (a drug information source 210) for ENBREL[®] brand etanercept includes the following instances of adverse event content 220 in table form.

[0058]

| <u>Percent of RA Patients Reporting Adverse Events ...</u> | | |
|--|------------------------------|-----------------------------|
| <u>Percent of patients</u> | | |
| <u>Event</u> | <u>Placebo (n = 152)</u> | <u>ENBREL (n = 349)</u> |
| Injection Site reaction | 10 | 37 |
| Infection | 32 | 35 |

[0059] Each instance of adverse event content 220 includes least one adverse event characterization 230. In the above example, the set {ENBREL, injection site reaction, 37%} is an adverse event characterization. This is an example of an adverse event characterization of the form { <drug in question> , <adverse event> , <frequency> }.

[0060] The drug label information for ENBREL[®] also includes the following adverse event content, in text form:

[0061] *ADVERSE REACTIONS ...*

[0062] *Other Adverse Reactions ...*

[0063] Other infrequent serious adverse events observed included: heart failure, myocardial infarction, ...

[0064] In this case, the adverse event characterization is quantitatively implicit and not quantitatively explicit, i.e., *heart attack* is characterized as *infrequent* as opposed to 1%. However, other information sources, such as accepted practice within the medical field or policy within a particular organization, may interpret *infrequent* as corresponding to a range of less than 2%. An exemplary characterization of *heart failure* with respect to *ENBREL*® in such a case would be {ENBREL, heart failure, infrequent} or {ENBREL, heart failure, < 2%}. Specific representations would be tailored to particular applications, e.g., < 2% could be represented by a range of 0% – 1.99%, i.e., a lower and upper range limit.

[0065] Preferred embodiments of the present invention include detecting instances of adverse event content from a drug information source. As with detection of drug rule content, this step can be accomplished manually or interactively in a computer-assisted manner. Natural language processing (NLP) is suited to providing computer assistance at this step. NLP processes keying on phrases such as "%" or keying on all sentences under a heading ADVERSE REACTIONS can suggest sections of text as candidate drug rule content. In some embodiments, the method described above is executed more than once as a validation, preferably involving interaction with different human users.

[0066] Referring to Figure 10, drug information sources 300 are typically associated with data about the information source, i.e., drug information source metadata, such as revision date or version. For example, the entry for Merrem® in the 2000 edition of PDR® indicates that the label version is "Rev E 3/99." Embodiments of the present invention extract 310 such drug information source metadata as one element to characterize the drug in a rationalized database 370.

[0067] Drug information sources may also typically contain descriptions of:

[0068] • circumstances that provide the basis for initiation of a treatment using the drug, i.e., *indications* ;

[0069] • symptoms or circumstance that renders the use of a drug inadvisable, i.e., *contraindications* ;

[0070] • factors a practitioner should consider when prescribing a drug, i.e., *precautions*

and *warnings*.

[0071] Given that verbatim terms referring to the same condition, compound, symptom, etc. may vary across labels (and even within a label), preferred embodiments of the present invention provide mapping 340 from verbatim drug information source 300 terms to a set of standard terms that will serve as a basis for query. The combination of verbatim terms data mapped to standard token terms serves as a thesaurus. In one embodiment, MedDRA™ serves as the set of standard terms 360. In other embodiments, a user may select the dictionary of reference terms 360.

[0072] Transparency in the process of moving from source data verbatim terms to a cleaned safety database with verbatim terms mapped to tokens is important to both database developers/operators and to end users. Preferred embodiments of the present invention capture the way source data terms have been cleaned and mapped as the "pedigree" of each term. The "pedigree" of a term is the link between the mapped term and the decisions made during data cleanup. End users typically wish to verify the pedigree of the data they use. In those embodiments, retained data includes one or more of the following as appropriate: verbatim term, token mapped to, source of the verbatim term, number of occurrences of the verbatim term, which type of cleanup (if any) was performed, and a cross-reference to where the token is defined.

[0073] Adverse event data, typically collected during clinical trials can be found in some drug information sources, e.g., drug labels, in both tabular and text form. Embodiments of the invention process 320 this data as described above, from the drug information sources 300 containing such data. These embodiments identify adverse event data in a manner amenable to query or use as input for an artificial intelligence engine. Drug rule data is also processed 350 in the embodiments illustrated in Figure 3; preferably in the manner described earlier for processing drug rule data.

[0074] In preferred embodiments illustrated in Figure 3, the full-text of the drug information source 300 (including graphs, tables, charts, pictograms) is associated 370 with the source 300 in the database. The set of adverse event data, drug rules, metadata, and full text serve to characterize the drug. The set of characterizations serves as the database to which various analytical engines (e.g., neural nets, case-

based reasoning tools, and predicate calculus engines) can be applied.

[0075] Preferred embodiments of the present invention include those implemented on a single computer or across a network of computers, e.g., a local area network of the Internet. Preferred embodiments include implementations on computer-readable media storing a computer program product performing one or more of the steps described herein. Such a computer program product contains modules implementing the steps as functions inter-related as described herein. Preferred embodiments of the invention include the unique data structures described herein, encoded on a computer-readable medium and computer signals transmissible over a computer/communications network.

[0076] A method and system for rationalizing drug label data has been described herein. These and other variations, which will be appreciated by those skilled in the art, are within the intended scope of this invention as claimed below. As previously stated, detailed embodiments of the present invention are disclosed herein; however, it is to be understood that the disclosed embodiments are merely exemplary of the invention that may be embodied in various forms.